FITS 变长数组在暗物质卫星数据存储中的应用研究

杨晓艳[1,2]，石涛[1,2]，李冰[1,2]，孙小涓[1,2]，卢晓军[3]

（1 中国科学院空间信息处理与应用系统技术重点实验室　北京　100190

2 中国科学院电子学研究所　北京　100190

3 中国国际工程咨询公司　北京　100048）

摘要：FITS 是空间天文领域广泛使用的一种数据格式，空间天文数据文件通常采用定长数据结构存储为 FITS 文件。作为我国首颗发射的空间科学卫星，暗物质粒子探测卫星科学数据源包具有长度可变的特点。在暗物质卫星数据处理过程中，急需设计支持可变长度数据的存储结构和处理方法。设计了一种支持长度可变数组的 FITS 格式，并实现了采用该数据结构的数据处理、存储和管理。应用于暗物质卫星地面处理中，验证结果表明，该方法实现了编辑级产品数据存储，显著降低了产品数据量，节约了存储空间，同时提升了处理效率。

关键词：FITS；变长数据结构；空间天文；暗物质卫星

中图分类号：P172.2　**文献标识码：**A　　文章编号:1672-7673(2018)

FITS（Flexible Image Transport system）①是一种在空间天文领域广泛使用的数据格式[1]，目的是为了传输、分析和归档天文科学数据文件。自 20 世纪 80 年代 FITS 格式被国际天文联合会（IAU）确认为国际标准以来，大部分天文数据以 FITS 格式在世界各地的数据中心存储和交换。

美国 Chandra②卫星、Swift③卫星、欧洲 EXOSAT④卫星等国际知名的天文卫星数据，欧洲低频射电干涉阵列（Low Frequency Array，LOFAR）⑤、澳大利亚望远镜致密阵列（Australia Telescope Compact Array，ATCA）[2]等地基天文观测数据，以及我国 HXMT 卫星[3]数据均采用 FITS 格式存储。以中国科学院国家天文台为首的中国天文学界联合建设了中国虚拟天文台（China-VO）[4]，针对系统中 FITS 文件检索与访问方面的问题，文[5-6]进行研究并构建了 FiHAS 系统。

通常情况下，空间天文数据采用定长数组的方式存储为 FITS 文件。FITS 格式文件由整数个长度为 2880 字节的报头和数据单元（Header and Data Unit,HDU）组成，其数据单元区采用长度固定的 ASCII 表或者二进制表存储。但是，暗物质粒子探测卫星的科学数据源包长度随有效载荷探测模式、粒子击中状态的不同而不同，定长数据的存储方式无法满足其数据存储的需要；同时，暗物质卫星是一颗空间天文卫星，出于数据共享的需求，其编辑级产品文件必须采用 FITS 格式存储。因此，需要根据暗物质卫星的数据特点，设计并实现一种数组长度可变的 FITS 格式数据存储方案。

基金项目：中国科学院空间战略性先导专项地面支撑系统数据处理与管理分系统研制项目资助.
收稿日期:2017-08-31;修订日期:2017-09-18
作者简介：杨晓艳，女，硕士，研究方向：空间天文数据处理.Email:yangxiaoyan@mail.ie.ac.cn

① http://fits.gsfc.nasa.gov/standard21b.html
② http://cda.harvard.edu/chaser/mainEntry.do
③ https://swift.gsfc.nasa.gov/cgi-bin/sdc/ql?
④ https://heasarc.gsfc.nasa.gov/docs/exosat/archive/exosat_archive.html
⑤ http://lofar.target.rug.nl/

1 FITS 数据格式

FITS 数据格式能够在国际天文领域得到广泛应用，其原因之一是其自描述性和灵活性。标的 FITS 文件由一个主 HDU 和一定数量的扩展 HDU 组成，每个 HDU 都包括头单元和数据单元两部分。其中，主 HDU 的头单元包含该文件对应的卫星名称、生产日期等总体描述信息，支持扩展定义，数据单元为空；扩展 HDU 的头单元包含本 HDU 数据起始结束时间、参考坐标系、各列数据类型等元数据信息，也支持扩展定义，数据单元中以 ASCII 表或二进制表的行列存储数据信息[⑥]。标准的 FITS 文件结构如表 1。

表 1 FITS 文件结构示意

Table 1 FITS File structure

HDU	Primary HDU	Extension HDU		…	Extension HDU	
内容	Header Unit（Primary）	Header Unit (Extension)	Data Unit	…	Header Unit (Extension)	Data Unit
大小	2880*N	2880*N	2880*N	…	2880*N	2880*N
类型	ASCII 码	ASCII 码	二进制表\ASCII 表	…	ASCII 码	二进制表\ASCII 表

数据单元以 ASCII 表或二进制表形式存储二维数组，每行长度固定，每列的数据类型需保持一致。FITS 格式支持的数据类型包括：逻辑型（L），bit 型（X），字节型（B），16 位整型（I），32 位整型（J），64 位整型（K），单精度浮点型（E），双精度浮点型（D）。

另外，FITS 格式的灵活性还支持对变长数组进行存储，存储方法为在文件头单元中定义一组特殊关键字，指定变长数组起始位置的偏移量和数据总长度。变长数组的实体数据并不存放在数据单元中，而是存放在数据单元之后的 heap 区域中。

2 暗物质卫星数据特点和存储要求

暗物质粒子探测卫星是中国科学院空间科学战略性先导科技专项中首批确定的五颗科学卫星之一，旨在通过高精度测量高能电子和伽玛射线能谱及其空间分布进行暗物质粒子探测，寻找和发现宇宙暗物质粒子，对其可能的宇宙空间分布进行研究。暗物质卫星已于 2015 年 12 月发射。

暗物质卫星有效载荷获取的科学观测数据以及卫星平台采集的工程数据通过数传通道下传至地面，经过帧同步、虚拟信道分离、源包提取、验证排序、排重、拼接/切分等处理后按照 FITS 格式输出为编辑级产品，完成产品归档存储，并分发给科学应用系统。主要处理流程如图 1。

⑥ https://heasarc.gsfc.nasa.gov/fitsio/fitsio.html

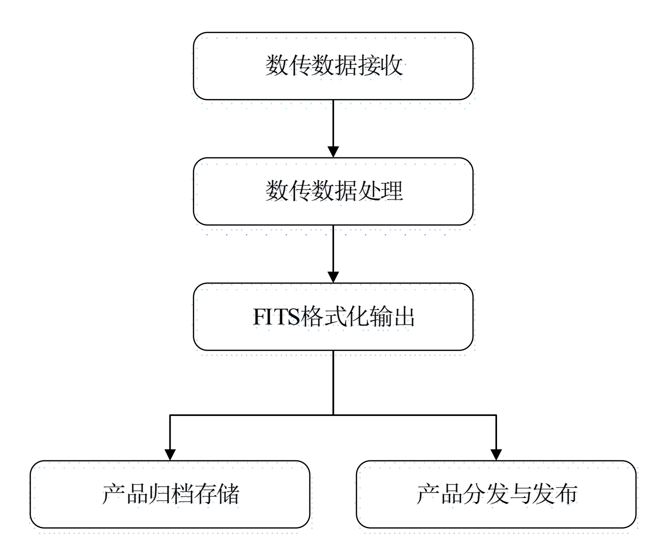


图 1 暗物质卫星数据处理、存储及分发流程

Fig. 1 data processing, storage and distribution processes of DAMPE

暗物质卫星数传数据中，卫星平台采集的表示载荷、平台工作状态的工程数据长度是固定的按照通用的定长方式存储即可；而科学观测数据源包由 30 个数据帧队列组成，总长度可变，源包结构以及各数据帧最大长度如表 2。

表 2 暗物质卫星科学数据源包结构

Table 2 The science data source package structure of DAMPE

序号	数据帧类型		最大数据帧长/bit
1	触发逻辑包		12
2	中子探测器数据帧		18
3	+*X* 方向塑闪 、Si、BGO 数据帧	+*X* 方向塑闪数据帧	178
4		+*X* 方向 Si FEE1 数据帧	2000
5		+*X* 方向 Si FEE2 数据帧	2000
6		+*X* 方向 BGO FEE1 数据帧	302
7		+*X* 方向 BGO FEE2 数据帧	302
8		+*X* 方向 BGO FEE3 数据帧	158
9		+*X* 方向 BGO FEE4 数据帧	302
10~16	-*X* 方向塑闪、Si、BGO 数据帧		5242
17~23	+*Y* 方向塑闪、Si、BGO 数据帧		5242
24~30	-*Y* 方向塑闪、Si、BGO 数据帧		5242
暗物质卫星科学数据源包总长度			**20998**

暗物质卫星每轨下传的数传数据中，科学数据源包数量为 60 万左右，按照每个源包 20998 字节的最大长度计算，单个文件大小为 11.73GB。但实际上，由于科学数据帧的实际长度与有效载荷模式、粒子击中状态有关，而且星上可能会对数据进行压缩，因此，科学数据源包长度是变化的。尤其是硅阵列探测器，总共有 7 万多个通道，绝大多数情况下，没有大击中事例发生，其大多数通道没有响应，并不输出科学数据。如果按照最大数据量存储，会造成暗物质卫星科学数据文件量偏大，导致数据处理和应用的难度增加、效率降低，以及存储资源的巨大浪费。

针对暗物质卫星科学数据源包特点及其存储需求，需要设计并实现一种能够支持变长

数据的 FITS 文件存储方案。

3 FITS 可变长存储方案设计与实现

如上所述，暗物质卫星科学数据源包长度变化范围比较大，以下针对其数据特点和存储需求，设计数据存储方案，并采用 C++语言调用 CFITSIO[12]完成软件实现。暗物质卫星数据存储方案设计如下：

（1）文件头改造。

在通用的定长数组存储方案中，文件头中用关键字 tform 指定数据类型，包括 L、X、B、I、J 等类型。在暗物质卫星变长数据存储方案中，tform 将赋值为 rPt(emax)或 rQt(emax)两种特殊类型。其中，*r* 为计数，可以是 0、1 或者不出现；P、Q 为数组描述符类型，分别表示 32 位、64 位有符号整数；t 为数据类型代码；emax 为数据长度最长的字节数。根据暗物质卫星数据类型，设置 tform=1QB（emax），表示数组描述符类型为 64 位有符号整数，实体数据按字节类型存储，emax 值从实际数据中提取。

另外，关键字 theap 表示 heap 区域的开始位置，省略时默认值为数据单元长度，表示 heap 区域直接从数据单元的下一个字节开始；如果取值大于默认值，表示 heap 区域与数据单元区域之间有一定的间隔。pcount 为间隔大小与 heap 区域大小之和。暗物质卫星变长数据存储方案中，theap 取默认值，pcount 为科学源包数据的字节数，该参数值从实际数据中提取。

（2）调用 fits_create_tbl 函数，创建 FITS 文件。

函数调用方式如下：

fits_create_tbl(fitsfile *fptr, int tbltype, LONGLONG naxis2, int tfields, char *ttype[],char *tform[], char *tunit[], char *extname, int *status)

其中，fptr 表示准备创建的暗物质卫星编辑级产品 FITS 文件；

tbltype 表示数据单元区表格类型，ASCII_TBL 表示 ASCII 表，BINARY_TBL 表示二进制表，暗物质卫星数据采用二进制表存储；

naxis2 表示暗物质卫星科学数据源包总行数，该参数从实际数据中提取；

tfields 表示参数个数，暗物质卫星科学数据源包产品仅有 CCSDS 源包 1 个参数列；

ttype 表示参数名称，命名为 CCSDS；

tform 表示参数类型，如上文所述，设置 tform=1QB(emax)；

tunit 表示参数度量单位，CCSDS 源包没有单位；

extname 表示扩展 HDU 的名称，命名为 Sci_Src。

（3）写入数组描述符。

数组描述符是一个 N*2 的二维矩阵，N 为变长数组的总行数，第 1 列定义数组中各行数据的长度，第 2 列定义各行数据起始位置在整个 heap 区域的偏移量，取值均为正整数，存储在数据单元中。暗物质卫星变长数据存储方案中，创建两个 N 维索引数组，分别命名为 index_len[N]，index_offset[N]，N 表示暗物质卫星科学数据源包总行数，从实际数据中提取每行长度，存入 index_len 数组，提取每行数据起始位置偏移量，存入 index_offset 数

组。（4）在 heap 区域中写入变长数组数据。

变长数组实体逐行存在 heap 区域中，由于 theap 取默认值，因此，存储起始位置就是数据单元结束符的下一个字节。然后，根据数组描述符取值，确认各行数据的起始位置和各行长度，调用 fits_write_col 函数，将暗物质卫星科学数据源包数据逐行写入 FITS 文件 heap 区域，函数调用方式如下：

int fits_write_col(fitsfile *fptr, int datatype, int colnum, LONGLONG firstrow, LONGLONG firstelem, LONGLONG nelements, DTYPE *array, > int *status)

其中，fptr 表示准备写入的暗物质卫星编辑级产品 FITS 文件；

datatype 表示写入方式，暗物质卫星数据按字节写入，该参数设置为 TBYTE；

colnum 表示行号，从第 1 行开始；

firstrow 表示起始写入的行号，从 1 开始；

firstelem 表示该行的第一个元素；

nelements 为该行数据长度，取值为 index_len[N]；

*array 为准备写入该行的 CCSDS 源包数据的存储位置，设置为 p+index.index[N].offset，其中 p 为位置指针。

按照上述方案，暗物质卫星数据 FITS 文件中扩展 HDU 的存储结构如下：

【文件头】	【数据单元】 存储暗物质卫星数据数组描述符	【heap】 存储暗物质卫星科学源包数据实体

4 效果验证

为了验证上述存储方案在产品文件数据量、存储效率、处理效率等方面的性能，选择 2016 年 4 月 29 日（2041 圈）、2016 年 6 月 5 日（2605 圈）、2017 年 7 月 3 日（8595 圈）、2017 年 7 月 25 日（8923 圈）、2017 年 8 月 30 日（9478 圈）共 5 轨暗物质卫星数传数据中科学源包类数据的编辑级产品文件进行分析。表 3 对定长 FITS 方案存储与变长方案存储的暗物质卫星编辑级产品文件大小进行了对比。

表 3 两种格式下 FITS 文件数据量对比

Table 3 Comparison of the two format file sizes

序号	源包数量（个）	源包最大长度（字节）	定长 FITS 格式文件大小（GB）	暗物质产品实际文件大小（GB）	数据量降低（%）
1	507797	20998	9.93	1.16	88.36
2	328236	20998	6.42	0.76	88.21
3	610706	20998	11.94	1.49	87.56
4	725230	20998	14.18	1.72	87.85
5	529467	20998	10.35	1.29	87.56

5 结果与讨论

真实数据的验证结果表明，针对暗物质卫星的数据产品存储特点，本文提出的基于FITS格式的变长数组存储方案能够将文件数据量降低88%左右，极大地节省了数据存储空间；同时，由于文件数据量的有效降低，数据的处理速度、产品的归档速度和应用效率都得到了明显提升。

本文提出的基于FITS格式的可变数组存储方案能够扩展应用到其他数据长度变化的天文数据存储中，该方案能够有效降低数据存储量，降低效率取决于实际数据长度与最大数据长度的平均比例。

参考文献

[1] Hanisch R J, Farris A, Greisen E, et al. Definition of the Flexible Image Transport System (FITS) [J]. Astronomy & Astrophysics, 2001, 376: 359-380.

[2] Murphy T，Lamb P，Owen C，et al. Data storage, processing, and visualization for the Australia telescope compact array [J]. Publications of the Astronomical Society of Australia，2006，23(1):25-32.

[3] 赵海升，葛明玉, 李正恒，等. 一种天文卫星数据预处理方法[J].天文研究与技术, 2017,14(6): 376-381.
Zhao Haisheng, Ge Mingyu, Li Zhengheng, et al. A method of data preprocessing for astronomical satellite [J] . Astronomical Research & Technology, 2017,14(6):376-381.

[4] 崔辰州，赵永恒. 中国虚拟天文台体系结构[J].天文研究与技术—国家天文台台刊，2004, 1(2)：140-151.
Cui Chenzhou, Zhao Yongheng. Architecture of Chinese Virtual Observatory[J]. Astronomical Research & Technology —Publications of National Astronomical Observatories of China, 2004, 1(2): 140-151.

[5] 崔辰州，李文, 于策，等. FITS数据文件的检索与访问[J].天文研究与技术—国家天文台台刊，2008,5(2)：116-123.
Cui Chenzhou, Li Wen, Yu Ce, et al, Search an location of FITS data files [J]. Astronomical Research & Technology —Publications of National Astronomical Observatories of China, 2008, 5(2): 116-123.

[6] 张海龙，冶鑫晨, 李慧娟，等. 天文数据检索与发布综述[J].天文研究与技术，2017,14(2):212-228.
Zhang Hailong, Ye Xinchen, Li Huijuan, et al. Astronomical data query and release review[J]. Astronomical Research & Technology, 2017,14(4):212-228.

Application Research of FITS Variable-Length Arrays in DAMPE Data Storage

Yang Xiaoyan[1,2], Shi Tao[1,2], Li Bing[1,2], Sun Xiaojuan[1,2] , Lu Xiaojun[3]

1. Key Laboratory of Technology in Geo-spatial Information Processing and Application System, Chinese Academy of Sciences, Beijing 100190, China

2. Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China

3. China International Engineering Consulting Corporation, Beijing 100048, China

Abstract: FITS (Flexible Image Transport system) is a data format widely used in space astronomy, and astronomical data files are usually stored in a fixed-length data structure as FITS files. As the first space science satellite in our country, DAMPE (Dark Matter Particle Explorer) satellite's science data source package has the characteristics of variable length. In DAMPE data processing process, it is urgent to design a storage structure and processing method that supports variable length data. This paper designed a FITS format that supports variable-length arrays and implemented DAMPE data processing, storage, and management with this data structure. The validation with real data of DAMPE indicated that this method was successfully applied to editing grade products storage of DAMPE, not only significantly reducing the amount of product data, saving storage space, but also improving the speed of data processing.

Keyword: FITS; Variable-length arrays; Astronomy; DAMPE